AIRS/AMSU/HSB Version 5 Level 3 Quantization Product Quick Start

Edited by:
Edward T. Olsen

Contributions by:
A. Braverman, S. Granger, E. Manning
Jet Propulsion Laboratory, California Institute of Technology

June, 2007
Version 1.0

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA

Submit Questions to:
http://airs.jpl.nasa.gov/AskAirs

Table of Contents

Table of Figures

Background

Members of the research community are increasingly interested in treating the Earth's weather and climate systems as nonlinear, suggesting that simple means and standard deviations are not sufficiently detailed representations for their studies. In an effort to address the needs of the community, the AIRS Project provides advanced Level 3 products containing cluster analyses, a method which is capable of capturing atmospheric variability in a potentially more representative manner than the AIRS Level 3 Standard Product.

Introduction

The V5 Level 3 Quantization (L3Q) Products are distributional summaries derived from the Level 2 standard swath products. The quantization approach compresses lower level calibrated and geolocated measurements at instrument resolution (Level 2) in a manner that provides a more comprehensive set of statistical summaries than simple means and standard deviations. These aggregates in space and time preserve the multivariate distributional features of the original data and so provide a compressed data set that more accurately describes the disparate atmospheric states that appear in the original high-volume data set.

The L3Q products combine the Level 2 standard data parameters over grid cells of 5° x 5° spatial extent for temporal periods of five days and one month. The report for each grid cell provides a number of representative values and their associated weights. The sample of lower level data points with the same representative is called a "**cluster**", and its representative is the "**cluster mean**". Cluster membership is determined by assigning lower level data points to clusters so that that the errors between original values and their representatives are minimized. This constraint prevents the creation of single-data-point clusters.

The resulting statistical summaries can be thought of as high-dimensional histograms, with the following properties:

- Cluster means are "centers-of-mass", not mid-points
- Cluster shapes are allowed to adapt to the shape of the data
- Cluster errors are the average squared Euclidian distance between the representatives (cluster means) and all members of the cluster
- Histograms representing different grid cells are comparable in their quality as representations of the underlying data

Figure 1 depicts the situation in the case where there are only two low-level variables rather than the 35 that are present in the AIRS L3Q product.

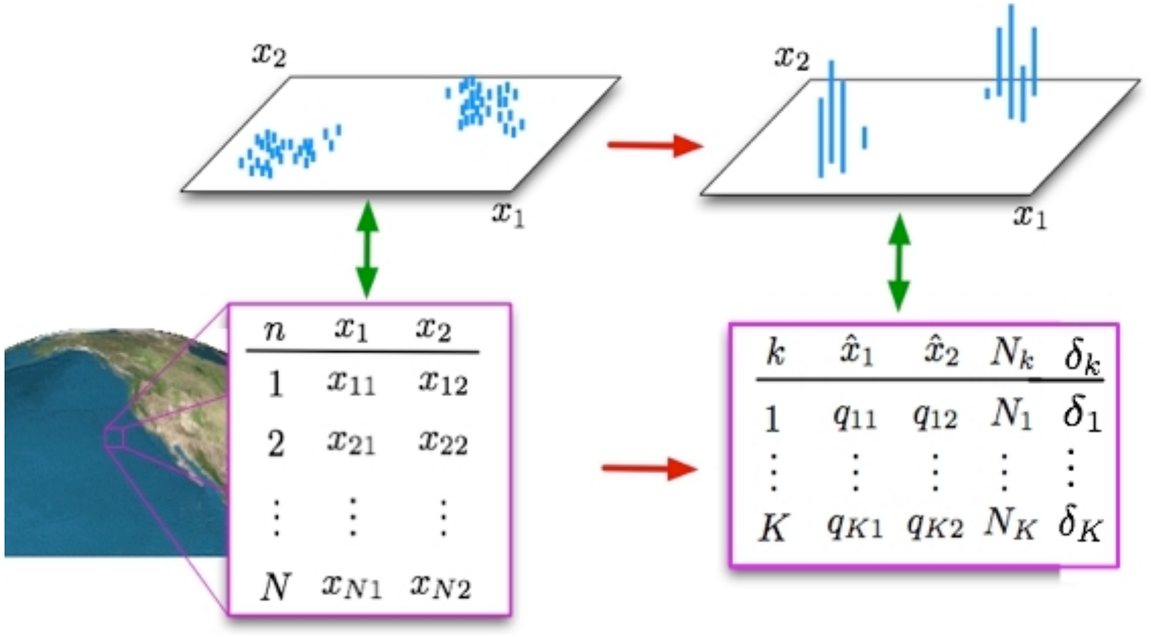

Figure 1: Schematic of AIRS Level 2 data for two variables in one grid cell and the corresponding cluster summarization.

The top left image in Figure 1 shows a scatter plot in which each Level 2 data point has equal mass and the bottom left image shows the corresponding data table. There are **N** Level 2 data vectors, and two variables, $\mathbf{x_1}$ and $\mathbf{x_2}$. The transformation into clusters is represented by the images on the right. The top right image shows the cluster representatives, each of which has a different "mass" corresponding to the cluster count, $\mathbf{N_k}$. The bottom right image shows the corresponding cluster table. **K** is the number of clusters and is smaller than the number of Level 2 data vectors, **N**.

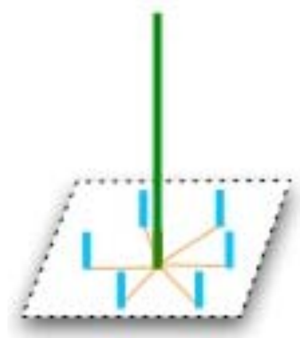

Figure 2: Zoom in to a single cluster representative

Figure 2 zooms in to the k-th cluster representative, showing the Level 2 data vectors assigned to that cluster. The distances between the representative and its members (in this example, two-dimensional and shown in gold) form the basis for calculating the cluster error. The cluster error, δ_k, is the average of the squared distances of the $\mathbf{N_k}$ Level 2 measurements that are represented:

$$\delta_k = \frac{1}{N_k}\sum_{n=1}^{N_k}\left\|\hat{x}_{nk} - x_{nk}\right\|^2$$

Each AIRS L3Q cluster may be thought of as a state (described by 35 parameters and their errors covering air temperature profile, specific humidity profile, cloud fraction profile and fraction of FOV that is land). The errors of each parameter are a measure of the variability of the samples that are included. The L3Q Products may be best understood as an extension of the traditional Level 3 Standard Products. If the L3Q Products were constrained to provide exactly one representative per grid cell they would provide the same information except for the fact that the individual variable counts and standard deviations would all be rolled into single values for count and error.

By allowing multiple representative vectors per grid cell, L3Q achieves two improvements over the traditional Level 3 Standard Products. First, it approximates the Level 2 data distribution. The mean and standard deviation only fully specify a data distribution if that distribution is Gaussian. Second, by treating AIRS data as vectors, it approximately preserves the joint relationships among variables. Joint relationships among pairs of variables are only captured if covariances or correlations are explicitly provided in traditional Level 3 products. Even if they are, these statistics are measures of linear association only. We have every reason to suspect that important relationships among the atmospheric state variables captured in the AIRS data may be non-linear.

The larger the number of representatives used for a grid cell, the better will be the approximation of the original Level 2 multivariate data distribution by the summary. In the extreme, if the number of clusters equals the number of raw data vectors then every cluster contains just a single AIRS Level 2 data vector, which is the cluster representative. The count in each cluster is unity and the errors (squared distances) are zero. This provides a perfect representation of the original data, but no data compression is achieved.

At the other extreme, allowing only one cluster per grid cell maximizes the data compression. Unfortunately the error is also maximized. The algorithm that creates the L3Q Product seeks a good compromise between these two extremes by employing information-theoretic tests to determine whether additional error is incurred as a result of increased compression of the data. The number of clusters (and the assignment of Level 2 data vectors to them) is modulated to reflect the information-theoretic complexity of the data being summarized. For example, in a grid cell where almost all the Level 2 data vectors are very nearly the same, only one cluster would be required to preserve most of the information with little error. However, an extremely heterogeneous grid cell requires a larger number of clusters to represent the multivariate data distribution with a similar low error.

The representatives and their weights may be used as proxies for the lower level (Level 2) data they summarize with the understanding that some information loss is suffered. That loss is quantified by the mean squared error between the summary and the original data, and is reported as part of the L3Q product.

As a general rule, Level 2 retrieved quantities whose quality indicators are “best” (=0) or “good” (=1) are included in the data that generate the L3Q product. Information contained within the two documents below will not be repeated in this document, and users **MUST** read them to avoid misuse of the L3Q products in their research.

V5_L2_Standard Product_QuickStart.pdf

V5_L2_Quality_Control_and_Error_Estimation.pdf

H2OpresLvls are the pressure levels upon which the temperature and moisture products are reported in the L3Q products. The values (in mb) are provided for convenient reference in the document:

V5_L3_Standard_Pressure_Levels.pdf.

Pentad and Monthly Products

The AIRS L3Q Product provides distributional summaries of 35 AIRS geophysical parameters with a spatial resolution of 5°x5° and temporal resolutions of 5 days (pentad) and a calendar month. It should be noted that the pentad product is constrained to fall within the calendar month and thus some pentad products may contain as few as 3 days of data or as much as 6 days. Pentads always start on the 1st, 6th, 11th, 16th, 21st, and 26th days of the month.

Example L3Q Product File Names

The following examples are L3Q pentad and monthly product files for December, 2009. File names include a date in the format of YYYY.MM.DD. YYYY is the year, MM is the month and DD is the day of month. This date is the beginning of the data period included in the pentad or monthly L3Q product.

Pentad Product Dec 3, 2009:

Name: AIRS.2009.12.03.L3.RetQuant005.v5.0.14.0.G2002123120634.hdf

Shortname: AIRX3QP5

Monthly Product Dec, 2009:

Name: AIRS.2009.12.01.L3.RetQuant031.v5.0.14.0.G2002123120634.hdf

Shortname: AIRX3QPM

L3Q Dimensions

The user should be aware of the following dimensions when referencing L3Q data parameters:

Dimensions		
Name	**Value**	**Description**
LonDim	72	Number of Longitude grid cells. 72 5-degree cells amount to 360 degrees. Cells are ordered West to East, from -180 to +180.
LatDim	36	Number of Latitude grid cells. 36 5-degree cells amount to 180 degrees. Cells are ordered North to South, from -90 to +90.
NumTrials	200	Number of different clustering attempts for each grid cell.
MaxNumClusters	100	Maximum number of clusters permitted in each grid cell. Actual number of clusters can be less. In this case, only the first **NumClusters** values are valid.
NumDimNorm	18	Dimensionality of clusters in normalized space.
NumDimPhysical	35	Dimensionality of clusters in physical space. See L3Q Products, below, for table of the products.
NumPentad	6	Present in monthly files only – Number of pentads contributing to month. (6 5-day periods gives 30 days. For longer or shorter months the last pentad will be 3-6 Days. See TBD.)

L3Q Product Attributes

L3Q attributes appear once per granule. They provide additional information that is common and therefore need not be replicated for each grid bin.

Global Attributes		
Name	**Additional Dimensions**	**Description**
Start_year	None	Year at start of data set
Start_month	None	Month at start of data set
Start_day	None	Day at start of data set. Data starts at the beginning of this day.
Start_TAI	None	TAI93 at start of data set
End_year	None	Year at end of data set
End_month	None	Month at end of data set
End_day	None	Day at end of data set. Data runs through the end of this day.
End_TAI	None	TAI93 at end of data set
Means	NumDimPhysical	Means of Physical Parameters (T, q…)
Covariance Matrix	NumDimPhysical, NumDimPhysical	
Eigenvectors	NumDimPhysical, NumDimPhysical	
PhysicalValuesDescriptor	NumDimPhysical strings	An array of string values describing the contents of PhysicalValues. (e.g., “Temperature at 350 mb (K)”)
Lambda	None	

L3Q Product Grids

Each L3Q Product (pentad and monthly) consists of a single grid containing fields created using the appropriate Level 2 products whose quality indicators are “best” or “good. The grid provides a 72x36 array of L3Q products.

Name	Additional Dimension	Description
LatCenter	None	Center Latitude of 5x5 grid cell (-90.0, +90.0), degrees North
LonCenter	None	Center Longitude of 5x5 grid cell (-180.0, +180.0), degrees East
SouthLatBound	None	Minimum bounding latitude in a 5x5 degree grid cell. (-90.0, +90.0), degrees North
NorthLatBound	None	Maximum bounding latitude in a 5x5 degree grid cell. (-90.0, +90.0), degrees North
WestLonBound	None	Minimum bounding longitude in a 5x5 degree grid cell. (-180.0, +180.0), degrees East
EastLonBound	None	Maximum bounding longitude in a 5x5 degree grid cell. (-180.0, +180.0), degrees East
NumClusters	None	Number of clusters in a 5x5 degree grid cell. Cannot exceed **MaxNumClusters**, unitless
Continued on next page	Continued	Continued on next page

Name	Additional Dimension	Description
NormalizedValues	MaxNumClusters, NumDimNorm	Normalized observations averaged over each cluster, unitless
PhysicalValues	NumClusters, NumDimPhysical	Raw physical observations averaged over each cluster. **PhysicalValuesDescription** in Global Attributes gives mapping of contents to physical values (e.g., T, H2O…), various physical units. See table of physical variables in following section, L3Q Products.
NumObsInCluster	MaxNumClusters	Number of Observations represented by this cluster, unitless
ClusterMeanSquaredError	MaxNumClusters	
Entropy	NumTrials	
GridCellMeanSquaredError	NumTrials	
PentadComposition	MaxNumClusters, NumPentad	Present in Monthly files only. Number of observations in each cluster derived from each pentad. Values must sum to **NumObsInCluster**, unitless
Cluster Distortion	MaxNumClusters	

L3Q Products

Variables Summarized by the AIRS Level 3 Quantization Product	
Index	**Physical Variable**
1	Atmospheric temperature at 150 mb level
2	Atmospheric temperature at 200 mb level
3	Atmospheric temperature at 250 mb level
4	Atmospheric temperature at 300 mb level
5	Atmospheric temperature at 400 mb level
6	Atmospheric temperature at 500 mb level
7	Atmospheric temperature at 600 mb level
8	Atmospheric temperature at 700 mb level
9	Atmospheric temperature at 850 mb level
10	Atmospheric temperature at 925 mb level
11	Atmospheric temperature at 1000 mb level
12	Specific humidity in layer bounded by TOA and 150 mb
13	Specific humidity in layer bounded by 150 mb and 200 mb
14	Specific humidity in layer bounded by 200 mb and 250 mb
15	Specific humidity in layer bounded by 250 mb 300 mb
16	Specific humidity in layer bounded by 300 mb and 400 mb
17	Specific humidity in layer bounded by 400 mb and 500 mb
18	Specific humidity in layer bounded by 500 mb and 600 mb
19	Specific humidity in layer bounded by 600 mb and 700 mb
20	Specific humidity in layer bounded by 700 mb and 850 mb
21	Specific humidity in layer bounded by 850 mb and 925 mb
22	Specific humidity in layer bounded by 925 mb and 1000 mb
continued	continued on next page

Index	Physical Variable
23	Cloud fraction with cloud top pressure ≤ 200 mb
24	Cloud fraction for which 200 mb < PCldTop ≤ 250 mb
25	Cloud fraction for which 250 mb < PCldTop ≤ 300 mb
26	Cloud fraction for which 300 mb < PCldTop ≤ 400 mb
27	Cloud fraction for which 400 mb < PCldTop ≤ 500 mb
28	Cloud fraction for which 500 mb < PCldTop ≤ 600 mb
29	Cloud fraction for which 600 mb < PCldTop ≤ 700 mb
30	Cloud fraction for which 700 mb < PCldTop ≤ 850 mb
31	Cloud fraction for which 850 mb < PCldTop ≤ 925 mb
32	Cloud fraction for which 925 mb < PCldTop ≤ 1000 mb
33	Scene land fraction
34	Fraction of good quality observations
35	Fraction of observations that are daytime

Example Filtering of Clusters for Research

The AIRS Team will provide examples in a future update.

Disclaimer and Caveats for L3Q Data Products

The user is advised to read the full disclaimer documentation for the V5 Data Products Release:

V5_Data_Disclaimer.pdf

Caveats

AIRS Team is continuing to evaluate the L3Q product and will advise as the experience base grows.